W. Klitz,[1] Ph.D.; R. Reynolds,[2] Ph.D.; J. Chen,[3] Ph.D.; and H. A. Erlich,[2] Ph.D.

Analysis of Genotype Frequencies and Interlocus Association for the PM, DQA1, and D1S80 Loci in Four Populations

REFERENCE: Klitz W, Reynolds R, Chen J, Erlich HA. Analysis of genotype frequencies and interlocus association for the PM, DQA1, and D1S80 loci in four populations. J Forensic Sci 2000;45(5):1009–1015.

ABSTRACT: Allele frequencies of the LDLR, HBGG, GYPA, D7S8, GC, DQA1, and D1S80 loci are presented and genotypes are analyzed for each of four ethnic groups: African Americans (n = 200), US Caucasians (n = 200), US Hispanics (n = 200), and Japanese (n = 89). Hardy-Weinberg genotypic proportions were observed in all but two of the 28 population-locus tests undertaken. Those two instances are attributable to type I statistical error. Gametic equilibrium among loci is an assumption invoked for application of the product rule to utilize the discriminatory power from two or more loci simultaneously. Two statistical methods, a genotype matching statistic and log-linear modeling, were used to evaluate gametic disequilibrium. The match statistic, comparing observed to expected likelihood of genotypic identity for seven loci among pairs of individuals within the database, revealed only one statistically significant deviation among 20 tests. As expected, the probability of match was generally lowest in the test on all ethnic groups combined, indicating that allele frequencies differ among ethnic groups for some of the loci. This was confirmed with the statistic θ to measure ethnic stratification, in which about 0.10 of the genetic variation is apportioned among the four ethnic groups for four of the structural loci (LDLR, HBGG, GC, and DQA1), while for GYPA, D7S8, and D1S80, variation is more uniformly distributed among ethnic groups. Log-linear modeling was also applied to the five PM loci. The most parsimonious log-linear model included only three higher order terms: the two-way interactions of three of the PM loci with ethnic group. These three instances (LDLR, HBGG, and GC) indicated differences in allele frequencies between ethnic groups. No two or higher way interaction (disequilibrium) was observed among loci. In summary, the assumptions of Hardy-Weinberg and gametic equilibrium that facilitate the use of the five PM loci, DQA1 and D1S80 in forensic applications are consistent with the allele and genotype frequencies observed in these populations.

KEYWORDS: forensic science, population genetics, LDLR, HBGG, GYPA, D7S8, GC, DQA1, D1S80, African Americans, US Caucasians, US Hispanics, Japanese, Hardy-Weinberg genotypic proportion, polymarkers, PM loci

We report observed allele frequencies for US Caucasian, African-American, US Hispanic, and Japanese populations for seven genetic markers: HLA DQA1, LDLR, GYPA, HBGG, D7S8, GC, and D1S80. The results of statistical analysis for Hardy-Weinberg equilibrium are also reported. In addition, two distinct statistical analyses of the seven loci and four populations were undertaken to test the hypothesis of gametic disequilibrium among loci and heterogeneity among populations. A method for probability of genotypic matching among individuals is applied to all seven loci, and log-linear analysis is applied to the five PM loci. Finally, we measure the apportionment of genetic diversity at each locus across ethnic groups with the statistic θ.

Materials and Methods

Population Database Samples

The 200 US Caucasian, 200 African American, and 200 US Hispanic extracted DNA samples used in this study were generously provided by Dr. Marcia Eisenberg (Roche Biomedical Laboratories, now Laboratory Corporation of America). A different set of samples from the same source have been previously used for HLA DQA1 typing (without subtyping allele 4) to generate population databases (1). The 89 Japanese samples were provided by Takehiko Sasazuki (Kyushu University, Fukuoka, Japan).

PCR Amplification and Typing Procedures

For amplification of the LDLR, GYPA, HBGG, D7S8, GC, and HLA DQA1 markers, 2–10 ng of DNA were added to the reaction mix and primer set provided in the AmpliType PM+DQA1 Kit (PE Biosystems, Foster City, CA) or comparable reagents (i.e., development lots). The samples were amplified and subsequently typed according to the manufacturer's protocol. The PM and HLA DQA1 types were read from the strips by two individuals independently.

For the D1S80 locus, 5–10 ng of DNA were added to the reaction mix and $MgCl_2$ solution provided in the AmpliFLP D1S80 Kit (PE Biosystems), or comparable reagents, and amplified according to the manufacturer's protocol. The samples were typed on silver-stained GeneAmp Detection Gels (PE Biosystems) according to the manufacturer's protocol.

Statistical Methods

When the number of alleles at a locus (k) was three or more, Hardy-Weinberg testing of genotypic ratios observed at each locus was carried out using the exact test of Guo and Thompson (2). The chi square test statistic was applied when k was two. When a test was significant, deviations from the expected values for individual genotypes were examined.

[1] School of Public Health, University of California, Berkeley, CA 94720, and Children's Hospital Oakland Research Institute, Oakland, CA.

[2] Department of Human Genetics, Roche Molecular Systems, Alameda, CA.

[3] Department of Integrative Biology, University of California, Berkeley, CA.

Received 26 March 1999; and in revised form 3 Aug. 1999; accepted 17 Nov. 1999.

The independence of alleles at different loci, gametic disequilibrium, was examined with two distinct statistical approaches: the empirical frequency of matching among multilocus genotypes, and with log-linear modeling.

Pairwise Matching Method

We use the genotypic matching approach of Risch and Devlin (3), who developed a probability of match method for characterizing the presence of linkage disequilibrium among two or more loci. This approach is closely related to a statistic developed by Maynard-Smith et al. (4). By utilizing a resampling strategy to estimate variance, the method permits a test of interlocus association, suitable for any number of loci and degree of polymorphism. It examines the probability of match for all pairs of individuals in the sampled population. We use the probability of match method to test the hypothesis of gametic disequilibrium between all pairs of the seven loci, and also for multiway associations between the PM loci, DQA1, and D1S80. This approach permits specific effects due to a particular locus or population to be discerned.

An observed value of the probability of match statistic is compared to the percentiles of the bootstrapped distribution to determine the statistical significance. The probability of match statistic is defined

$$Ts = (O(M) - E(M))^2/E(M)$$

where $O(M)$ is the observed number of matches in a sample of multilocus genotypes. $E(M)$ is the expected probability of match, which is the product of the observed genotype frequencies of the loci taken individually.

The bootstrap test is constructed as follows. Using the observed genotype frequencies for a locus, N individuals are randomly sampled with replacement from a population of size N. Each individual should have an equal probability of being selected, (i.e., $p = 1/N$). This step is performed for each of the loci in the set to be tested. The number of matches present for each of the genotypes formed from the loci in the set are calculated. This is the observed number of bootstrap matches, $OB(M)$. Compute the expected number of bootstrap matches, $EB(M)$, which is the product of the frequencies of genotypes calculated for each locus in the set times the sample size N. Calculate the bootstrap statistic, $Tb = (OB(M) - EB(M))^2/EB(M)$, and save the result. These steps are repeated a total of 1000 times, then the Tb samples are sorted by size. The percentiles of the bootstrapped samples are then compared with the observed Ts to determine the significance of the observed value.

Log-Linear Model

The log-linear model has been generalized for the analysis of contingency tables (5,6). Because of estimation problems with the large numbers of cells present in models which include the highly polymorphic loci DQA1 and D1S80, we apply log-linear modeling only to the five PM loci. The five-locus PM data can be displayed as a 6-dimensional contingency table of the following size: 3 by 3 by 6 by 3 by 6 by 4, which corresponds to the genotypes of the five loci as five dimensions and ethnic group as the other. There are 3888 cells for a contingency table of this size, but in the data set there are only 485 non-empty cells. Some of the genotype combinations never occurred in any of the four ethnic groups, while other combinations occurred in only some ethnic groups but not the others. As the patterns of occurrence of these empty cells (missing genotypes) seem intrinsic to the data, we decided to treat them differently. Those genotypes that never occur in the study are regarded as structural zeros and removed from the analysis. For those multilocus genotypes that occur only in some of the ethnic groups, the zero cells for the other ethnic groups are treated as sampling zeros and are included in the analysis (7). After these adjustments, the final table contains $363 \times 4 = 1452$ cells. To facilitate the model fitting, a small positive value ($1/e^8$) was added to cells having sampling zeros.

A full log-linear model includes terms for the grand mean, the first order terms and all terms for the two way and higher order interactions. A full log-linear model for the adjusted PM data contains one term for the mean, 6 first order terms, 15 second order terms, 20 third order terms, 15 fourth order terms, 6 fifth order terms and 1 sixth order term. This is a saturated model, and so is a perfect fit to the data. In order to identify which terms make meaningful contributions, simpler models are fitted. The simpler models are constructed from the full model by excluding terms, such as higher order interaction terms, or terms related to one particular locus, etc. For example, a pairwise interaction model contains only the average, the individual effects and two-way interaction terms. It has the following form:

$$\begin{gathered}\log (F_{ijklmn}) = \\ \mu + \\ \lambda_i^L + \lambda_j^G + \lambda_k^H + \lambda_l^D + \lambda_m^C + \lambda_n^E + \\ \lambda_{ij}^{LG} + \lambda_{ik}^{LH} + \lambda_{il}^{LD} + \lambda_{im}^{LC} + \lambda_{in}^{LE} + \\ \lambda_{jk}^{GH} + \lambda_{jl}^{GD} + \lambda_{jm}^{GC} + \lambda_{jn}^{GE} + \\ \lambda_{kl}^{HD} + \lambda_{km}^{HC} + \lambda_{kn}^{HE} + \\ \lambda_{lm}^{DC} + \lambda_{ln}^{DE} + \\ \lambda_{mn}^{CE}\end{gathered}$$

where F_{ijklmn} is the expected count for a particular cell having genotype *ijklm* and ethnic group n, μ is the overall mean, L, G, H, D, and C represent effects for the five PM loci (LDLR, GYPA, HGBB, D7S8, and GC, respectively), and E is the effect for ethnic variation. For example, λj^G is the effect of the jth genotype of the locus GYPA.

The fit of a simpler model can be compared with the full model by a likelihood ratio test. Let L_1 and L_0 be the log-likelihood functions of the restricted model (e.g., the pairwise interaction model) and a model that includes additional higher-order interaction parameters (e.g., the full model), respectively. Then

$$-2(L_1 - L_O) \sim \text{chi}^2_{\text{d.f.}}$$

where the number of d.f. for the chi^2 equals the difference of the numbers of parameters in the two models. If the model fitting is satisfactory, more terms are removed. If not, terms are added back to the model until the most parsimonious model is identified.

Inter-Population Variance (θ)

The degree of human population stratification on the loci examined was measured with θ (8), which, under the assumption of Hardy-Weinberg equilibrium and random mating, is equivalent to Wright's Fst statistic

$$\theta = (Ht - Hs)/Ht$$

where Ht is the heterozygosity in the entire combined population and Hs is the heterozygosity in each individual population.

Results

Allele Frequencies

A total of 689 samples from four population groups were typed: US Caucasians ($n = 200$), African Americans ($n = 200$), US Hispanics ($n = 200$), and Japanese ($n = 89$). The distributions of observed allele frequencies for the five PM markers HLA DQA1 and D1S80 are shown in Tables 1 and 2. The complete tables of allele frequencies and observed and expected genotype frequencies are available on request from Dr. Rebecca Reynolds (Rebecca.Reynolds@Roche.com).

Allele frequencies at the seven loci for each of the four populations were compared with population samples from the literature (9–15). The 28 tested comparisons (Refs 11 and 12) were nonsignificant (data not shown), except for the GC locus in Hispanics which differed from a reference sample (both Southwestern and Southeastern Hispanic) (12) at $p < 0.01$. If not attributable to type 1 statistical error, this result may be due to the heterogeneity of the category "Hispanic" as a group designation, and to the differentiation of the GC locus among ethnic groups (see below). In addition, allele frequencies at the LDLR locus differed from a reported New Jersey African-American sample (14) and at the D7S8 locus from a New Jersey Hispanic sample (14).

Testing Genotypic Ratios for Hardy-Weinberg Equilibrium

The Hardy Weinberg test for deviation from the genotypic ratios expected under random mating and in the absence of selection is a useful means of evaluating the quality of a data set. Any deviations observed might indicate hidden ethnic stratification of a population sample or typing errors. Significance testing results from HW exact testing for each of the 28 locus/population tests revealed that only two tests demonstrate nominally significant departures from HW expectations, HBGG in Caucasians at $p = 0.010$ and D1S80 at $p = 0.002$ in Hispanics.

A closer examination of the deviant cases reveals no evidence for population substructure, hidden ethnic, or systematic typing errors. In the Caucasian HBGG sample (three alleles, six possible genotypes; Table 3*A*), it is the observed frequencies of the genotypes containing the uncommon allele C that depart from expectations. Rare alleles can be the cause of significant departures from expectation, even if the population from which the samples are drawn is in Hardy-Weinberg proportion (8). Five copies of allele C are present in the Caucasian population sample (freq. = 0.0125), two of which comprise the one observed homozygote C/C. In addition, no heterozygotes of C with the most common allele A are observed. These two cells of the table are responsible for the deviation from HW expectation. If, for example, the single C/C

TABLE 2—*D1S80 allele frequencies in four populations.*

	Allele	Cauc.	Af Amer.	Hisp.	Japan
1	14			0.003	
2	16			0.003	0.045
3	17		0.048	0.013	0.006
4	18	0.238	0.098	0.263	0.146
5	19	0.010	0.003	0.005	0.017
6	20	0.040	0.033	0.020	
7	21	0.018	0.115	0.025	0.017
8	22	0.030	0.088	0.028	0.022
9	23	0.008	0.023	0.003	
10	24	0.348	0.193	0.318	0.219
11	25	0.040	0.023	0.055	0.011
12	26	0.015	0.008	0.010	0.006
13	27	0.013	0.013	0.008	0.034
14	28	0.063	0.153	0.050	0.090
15	29	0.053	0.055	0.055	0.051
16	30	0.008	0.008	0.055	0.146
17	31	0.080	0.048	0.058	0.124
18	32	0.013	0.005	0.003	
19	33	0.003	0.005		0.017
20	34	0.003	0.073	0.008	0.011
21	35	0.003			
22	36	0.005	0.003		
23	37	0.008		0.003	
24	38			0.005	
25	39				0.017
26	40	0.003	0.003	0.010	
27	41				0.006
28	>41	0.005	0.010	0.005	0.017

TABLE 1—*Allele frequencies for five PM loci and HLA DQA1 in four populations.*

Locus	Allele	Caucasian ($n = 200$)	African American ($n = 200$)	Hispanic ($n = 200$)	Japanese ($n = 89$)
LDLR	A	0.448	0.235	0.485	0.202
	B	0.553	0.765	0.515	0.798
GYPA	A	0.530	0.528	0.615	0.517
	B	0.470	0.473	0.385	0.483
HBGG	A	0.538	0.440	0.375	0.331
	B	0.450	0.228	0.580	0.669
	C	0.013	0.333	0.045	0.000
D7S8	A	0.610	0.655	0.623	0.612
	B	0.390	0.345	0.378	0.388
GC	A	0.275	0.090	0.203	0.287
	B	0.178	0.720	0.335	0.472
	C	0.548	0.190	0.463	0.242
HLA	1.1	0.158	0.125	0.105	0.084
DQA1	1.2	0.190	0.330	0.130	0.118
	1.3	0.073	0.058	0.053	0.236
	2	0.145	0.130	0.115	0.006
	3	0.193	0.090	0.218	0.444
	4.1	0.215	0.185	0.270	0.073
	4.2/4.3	0.028	0.083	0.110	0.039

TABLE 3—*Genotype frequencies in the two nominally deviant cases of Hardy Weinberg testing:* A. *HBGG genotypes in Caucasians,* N = *200 (HW exact* p = *0.010).* B. *Observed and expected values for deviant genotype frequencies of D1S80 in the Hispanic sample.*

	A	B	C	
A.				
A	59			
B	97	40		
C	0	3	1	
B.				
Genotype	17/17	18/18	18/24	24/24
Observed	2	9	53	15
Expected	0.01	13.8	33.3	20.2

homozygote were actually A/C, then the HW exact test *p* value would be 0.55. Some standard statistical methods for evaluating HW equilibrium (16) pool rare genotypes; using such methods, no deviation from HW equilibrium expectations for HBGG is found in this data set.

For the Hispanic D1S80 sample, the genotypes of the two most common alleles (18 and 24) and secondly, of the homozygotes of a rare allele (allele 17) all contribute to the deviation from HW expectation (Table 3*B*). Because typing for the four ethnic samples was carried out under identical conditions for each of the four population samples, a consistent typing error for the genotypes of alleles 18 and 24 seems an unlikely explanation. This is because both alleles 18 and 24 are common in the other three populations studied, and none of these other population samples have deviations for the genotypes 18/18, 18/24, or 24/24, as seen in the Hispanics (data not shown). The ethnic category "Hispanic" is notoriously poorly defined, including individuals whose genetic heritage is Native American, African as well as Iberian. In any case, the effect of population stratification is to reduce heterozygote frequencies, the opposite of the deviation reported here. It seems very unlikely that selection acting on the genotypes of this particular minisatellite could explain the deviant genotypic ratios.

In conclusion, in 7% (2/28) of the total tests, nominally significant deviations from Hardy Weinberg expectations were present. Thus, the observed deviations in the 28 tests do not represent an unusual outcome and are consistent with type I statistical error.

Tests of Gametic Disequilibrium

We take two approaches to examine interlocus associations among the five PM loci, DQA1, and D1S80. The absence of interlocus interactions would validate the application of the product rule with multilocus data for calculating probability of match statistics in forensic applications.

The first approach we present for evaluating gametic equilibrium among loci in multilocus genotype data is to examine the occurrence of matches within the sampled population (3,8). This approach provides an empirical test of the product rule for combining data from two or more loci as well as an estimate of the matching probabilities, which are of immediate forensic interest. Single locus genotypic matching results are shown in Table 4 for each of the five PM loci, and for DQA1 and D1S80 on each of the four population samples and for the combined data set. The number of paired comparisons in a sample of size *N* is $N(N - 1)/2$. In the case of the 200 Caucasians typed for LDLR, for example, 19,900 individual paired comparisons are present. The three genotypes of LDLR had 7367 matches, giving a probability of 0.37. Individually, the prob-

TABLE 4—*Single locus matching results. N, sample size; NC, number of comparisons; O(M), observed number of matches; P(M), probability of matches.*

				Polymarkers						
Population		N	NC	LDLR	GYPA	HBGG	D7S8	GC	DQA1	D1S80
African	O(M)	200	19,900	9,441	7,941	3,791	7,753	7,267	1,194	374
	P(M)			0.4744	0.3991	0.1905	0.3896	0.3652	0.0600	0.0188
Caucasian	O(M)	200	19,900	7,367	7,436	7,150	7,331	4,546	1,084	1153
	P(M)			0.3702	0.3737	0.3593	0.3684	0.2284	0.0545	0.0579
Hispanic	O(M)	200	19,900	7,221	7,832	6,504	7,419	4,361	1,005	1621
	P(M)			0.3629	0.3936	0.3268	0.3728	0.2192	0.0505	0.0815
Japanese	O(M)	89	3,916	1,982	1,588	1,522	1,416	767	442	96
	P(M)			0.5061	0.4055	0.3887	0.3616	0.1959	0.1129	0.0245
Combined	O(M)	689	237,016	91,241	92,232	58,801	89,346	45,647	11,190	8732
	P(M)			0.3850	0.3891	0.2481	0.3770	0.1926	0.0472	0.0368

ability of genotypic matches at the five polymarker loci is relatively high, ranging from a low value of 0.19 for African Americans at HBGG to a high of 0.51 for the Japanese at LDLR. Strong population specific effects on the probability of matching are absent, although we note that the Japanese sample does have the highest probability of match for four of the seven loci. DQA1 and D1S80 with their higher heterozygosities have probabilities of match ranging from 0.05 to 0.11 among the four populations. When all four populations are combined and tested for matching, probabilities of pairwise matches are at or near the bottom of the range of P(M) values present for the individual ethnic samples except in the case of D1S80. This implies that, in these cases, differences in allele frequencies among the ethnic groups tend to make the combined genotypic distributions more even, and so reduce the probability of matching in the "combined" group.

The multiplication of the match probabilities for the individual loci supplies an expected match probability that can be compared to the observed number of matches in multilocus comparisons for a test of gametic disequilibrium. We present results of pairwise matching on combinations of the seven loci (see Table 5). The expected probabilities of matching for the polymarker loci alone are approximately 3–6/1000 across the four populations. The addition of either DQA1 or D1S80 to PM reduces the expected match probability to the range 1–6/10,000 across the four ethnic groups. The genotypes of all loci examined together gives reduced probabilities of matching of 1–2/100,000. For the combined sample of all four populations, PM and PM with either DQA1 and D1S80 the probability of match is at the lower end of the range of values for the four populations examined separately. DQA1 and D1S80 genotypic matching probability is an order of magnitude less than that seen for the populations examined individually. The smallest probability of match expected is found when testing for all populations at all loci (5×10^{-6}).

If interlocus associations were present, then the observed numbers of matches would be greater than that expected based on individual locus frequencies. We note from Table 5 that for the five PM loci and for the other combinations of loci presented, that deviations from the expected number of matches fall in both directions. A consistent pattern of deviations is not present, either across populations or for the combinations of loci compared. In most instances the observed and the expected values are very similar. For a more formal statistical evaluation of the presence of gametic equilibrium and of the applicability of the product rule, we use the bootstrapped distributions to evaluate the probability of the match statistic Ts. Of the 20 tests only one—the six locus test of PM + DQA1 in Caucasians ($p < 0.05$)—is nominally significant (Table 5). We conclude that interlocus genotypic associations are absent.

A second approach for uncovering interlocus associations is log-linear modeling, a standard statistical method that has not been previously applied to forensics databases. To get a sense of possible pairwise interlocus effects, we first present the results for simple two way testing of each of the ten possible two-locus combinations of the five PM loci on each of the four populations, making a total of 40 tests (Table 6). Nominally significant two locus interactions are present in four instances: GYPA by GC in African Americans and Hispanics, LDLR by GYPA in Japanese and D7S8 by GC in Japanese. The *p*-values for three of these cases are close to 0.05. When the four populations are combined into a single sample and the ten two-locus tests are rerun, the three tests involving the loci LDLR, HBGG, and GC have high X^2 values. This result points to the possibility of locus by ethnic group interaction.

Results of the selective model fitting strategy for constructing the final log-linear model with consideration of all four ethnic groups and five loci are shown in Table 6. The pairwise interaction model, in which the results of selected model fitting, when all five PM loci are considered, are summarized in Table 7. The pairwise interaction model, in which all three-way or higher terms are removed from the full model, fits the data very well. After one further reduces the model by including only pairwise interaction terms related to the ethnic group variable, i.e., no interactions among the PM loci, (the population interaction only model), the fit is still extremely good ($p = 1$). This suggests that there is no evidence of interaction among these five loci, and the product rule for the calculation of matching probability could be applied to this data set. As

TABLE 5—*Multiple locus matching results. O(M), observed number of matches; E(M), expected number of matches; P(M), probability of matches; Ts, test statistic.*

Population	Loci Tested #Loci	PM 5	PM, DQA1 6	PM, D1S80 6	PM, DQA1, D1S80 7
African American	O(M)	97	7	0	0
	E(M)	102	6.1	1.9	0.1
	P(M)	0.00514	0.00031	0.00010	0.00001
	Ts	0.26	0.12	1.92	0.12
Caucasian	O(M)	99	9	7	0
	E(M)	84	4.6	4.5	0.3
	P(M)	0.00422	0.00023	0.00024	0.00001
	Ts	2.99	4.40*	0.98	0.26
Hispanic	O(M)	77	4	6	0
	E(M)	76	3.8	6.2	0.3
	P(M)	0.00381	0.00019	0.00031	0.00002
	Ts	0.02	0.01	0.01	0.31
Japanese	O(M)	28	4	0	0
	E(M)	22	2.5	0.5	0.1
	P(M)	0.00565	0.00064	0.00014	0.00002
	Ts	1.56	0.90	0.54	0.06
Combined	O(M)	645	36	27	0
	E(M)	640	30.8	23.5	1.1
	P(M)	0.00270	0.00013	0.00010	0.0000047
	Ts	0.05	1.12	0.50	1.11

* $p < 0.05$.

TABLE 6—*Pairwise interactions among the five PM loci for the four ethnic groups singly and combined using log-linear modeling. Nominally significant values are in bold.*

Pairwise Interaction		Ethnic Groups				
Loci	Statistic	African American	Caucasian	Hispanic	Japanese	Combined
LDLR, HBGG	X^2	12.26	10.36	7.89	3.40	**28.11**
	P-value	0.20	0.33	0.55	0.95	**0.001**
LDLR, GYPA	X^2	5.23	2.43	6.30	**10.52**	7.60
	P-value	0.27	0.66	0.18	**0.04**	0.11
LDLR, D7S8	X^2	1.79	4.89	1.37	0.82	3.74
	P-value	0.78	0.30	0.85	0.94	0.45
LDLR, GC	X^2	14.7	11.86	8.79	9.56	**46.50**
	P-value	0.15	0.30	0.56	0.49	**0.0001**
HBGG, GYPA	X^2	10.17	6.10	14.85	3.29	9.40
	P-value	0.43	0.81	0.14	0.98	0.50
HBGG, D7S8	X^2	6.23	5.71	12.73	3.14	3.77
	P-value	0.80	0.84	0.24	0.98	0.96
HBGG, GC	X^2	23.32	23.16	27.68	7.28	**90.11**
	P-value	0.51	0.52	0.28	0.99	**0.0001**
GYPA, D7S8	X^2	4.94	5.04	2.79	8.12	8.43
	P-value	0.30	0.29	0.60	0.09	0.08
GYPA, GC	X^2	**18.65**	8.06	**29.32**	16.38	14.96
	P-value	**0.05**	0.63	**0.002**	0.09	0.14
D7S8, GC	X^2	15.55	12.92	16.50	**20.33**	4.02
	P-value	0.12	0.23	0.09	**0.03**	0.95

TABLE 7—*Log-linear models tested on the PM data. The model best fitting the data, designated "Final Model," has locus by population interaction terms, but no locus-by-locus interaction terms.*

Model	Deviance	d.f.	*P*-Value
Full	0	0	0
Pairwise interaction	888.38	1289	1
Population interactions only	957.12	1384	1
Final Model: population by locus (GC, LDLR, HBGG) interactions only	967.70	1396	1
No interaction	1526.92	1432	0.04
No population	1588.63	1435	0.0027

the interactions between the ethnic group variable and loci D7S8 and GYPA are not statistically significant, we removed these two interactions from the model (the population and locus, LDLR, HBGG, GC, interaction only model), and find that it still fits ($p = 1$). When all pairwise interactions are excluded, the "no interaction model" no longer fits the data ($p = 0.04$). As expected, ethnic group interacts with the genotypic frequencies of some of the loci, i.e., locus GC, locus LDLR, and locus HBGG. When, for demonstration purposes, we further reduce the model by combining the data over all ethnic groups, i.e., the no population model, the fit gets much worse ($p = 0.0027$).

Based on the above analysis, we choose the population and locus (LDLR, HBGG, and GC) interaction only model as the final model. It has the following form

$$\log (F_{ijklmn}) =$$

$$\mu +$$

$$\lambda_i^L + \lambda_j^G + \lambda_k^H + \lambda_l^D + \lambda_m^C + \lambda_n^E +$$

$$\lambda_{kn}^{HE} + \lambda_{jn}^{LE} + \lambda_{mn}^{CE}$$

Although some significant pairwise locus interactions are present, even after controlling for ethnic group, these interactions disappear when all five PM loci are included in the model. Treating the data as a six-way contingency table, a final model is selected which contains only three pairwise interaction terms involving the ethnic group variable. The analysis shows that different ethnic groups have different profiles for the three PM loci, LDLR, HBGG, and GC. There is no evidence that the five PM loci are significantly associated with one another.

Population Variance and Ethnic Stratification Measured with θ

The second Committee on DNA Forensic Science report (8) has recommended the use of *θ*, a parameter that reflects population subdivision, to accommodate the possible effects of substructure in the calculations of estimated match probabilities. Thus, this approach does not assume Hardy-Weinberg proportions but uses procedures "that take deviations from HW into account" (p. 104). The value of *θ* for a given marker can be calculated from population data for different subgroups within an ethnic population. The Committee on DNA Forensic Science (8) states that, for PCR markers, a *θ* of 0.01 "would be appropriate" or "a more conservative value of 0.03 may be chosen" (p. 119). The data presented here are not derived from subpopulations so that a relevant *θ* for *within* population variation cannot be calculated. A value for *θ* that represents differences *between* ethnic populations, however, can be estimated.

The amount of genetic variation differentiating the four ethnic groups, *θ*, is presented for each locus in Table 8. Four of the five loci with expressed variation have moderately high levels of ethnic differentiation, near 0.10%. Of these loci GYPA stands out with a *θ* value of only 0.024. The two non-coding marker loci D7S8 and D1S80 have low values of *θ* at 0.001 and 0.019, respectively. These results correspond to data previously published (9–15), showing that little overall ethnic stratification occurs at these loci.

TABLE 8—*The apportionment of genetic variation between individuals and among ethnic groups for each locus.*

Locus	Inter-Individual Variation	Inter-Ethnic Group Variation (The Statistic θ)
LDLR	0.904	0.096
GYPA	0.976	0.024
HBGG	0.882	0.118
D7S8	0.999	0.001
GC	0.899	0.101
DQA1	0.909	0.091
D1S80	0.981	0.019

Discussion

We have analyzed the population characteristics of seven genetic loci (the five PM loci, HLA DQA1, and D1S80) for their appropriateness in forensic applications of individualization and for the use of the "product rule" in estimating multi-locus genotype frequencies. Consistent levels of heterozygosity are present at each of the loci in three major U.S. census groups (African American, Caucasian, and Hispanic) and the Japanese. In principle, the test of Hardy-Weinberg genotypic proportions can reveal systematic typing errors and population substructure. The genotypic ratios at nearly all loci and populations are within Hardy-Weinberg expectations. The two exceptions can be attributed to expected statistical type I error. These data indicate that, for the available power of the test, population substructure within any of the four populations and systematic typing errors are absent.

Combining the discriminatory power of two or more individual loci by multiplying the observed genotypic frequencies across loci (i.e., use of the product rule) assumes statistical independence among loci. Our two assessments of linkage disequilibrium among the seven loci uncovered no evidence of non-random distribution of alleles across loci. The empirically based resampling technique of pairwise genotype matching gave no evidence of two-way or higher order association among the seven loci. Similarly, the standard statistical treatment of multivariable categorical data, log-linear modeling, performed on the PM loci revealed the complete absence of higher order locus-locus interactions. Taken as a whole, the data and analyses presented here, affirm and validate the utility of these widely used markers in forensic science and allow the combination of genotype frequency data to determine multilocus genotype frequencies. Given the absence of statistical dependence, the combined use of these genetic markers can provide valuable probability of match statistics for forensic use.

Acknowledgments

This work supported in part by National Institutes of Health Grant GM35326 (W. Klitz, J. Chen). We acknowledge Mark P. Nelson's valuable programming contribution. The authors wish to thank N. Fildes, S. Cosso, J. Varlaro, M. Alavaren, and M. Maramba for their contributions to PM and D1S80 assay development and for typing samples.

References

1. Reynolds R, Sensabaugh G, Blake E. Analysis of genetic markers in forensic data samples using the polymerase chain reaction. Analytical Chemistry 1991;63(1):2–15.
2. Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 1992;48:361–72.
3. Risch NJ, Devlin B. On the probability of matching DNA fingerprints. Science 1992;255:717–20.
4. Maynard-Smith J, Smith NH, O'Rourke M, Spratt BG. How clonal are bacteria. Proc Natl Acad Sci USA 1993;90:4384–8.
5. Goodman LA. The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries. J Amer Statist Assoc 1968;63:1091–131.
6. Goodman LA. The multivariate analysis of qualitative data: interaction among multiple classifications. J Amer Statist Assoc 1970;65:226–56.
7. Agresti A. 1990, Categorical data analysis. New York: Wiley.
8. Committee on DNA Forensic Science, 1996. The evaluation of forensic DNA evidence. Commission on DNA Forensic Science: an update, National Research Council, National Academy Press.
9. Iminish, et al. Allele and haplotype frequencies for HLA and complement loci in various ethnic groups. In: Tsuji T, Izawa M, Sasazuki T, editors. Oxford University Press 1992;1:1065–220.
10. Rivas F, Zhong Y, Olivares N, Cerda-Flores RM, Chakraborty R. Worldwide genetic diversity at the HLA-DQA1 locus. Am J Hum Biol 1997;9:735–49.
11. Budowle B, Baechtel S, Smerick JB, Presley KW, Giusti AM, Parsons G, et al. D1S80 population data in African Americans, Caucasians, Southeastern Hispanics, Southwestern Hispanics, and Orientals. J Forensic Sci 1995;40(1):38–44.
12. Budowle B, Lindsey JA, DeCou JA, Koons BW, Giusti AM, Comey CT. Validation and population studies of the loci LDLR, GYPA, HBGG, D7S8, and Gc (PM loci), and HLA-DQα using a multiplex amplification and typing procedure. J Forensic Sci 1995;40(1):45–54.
13. Watanabe Y, Yamada S, Nagai A, Takayama T, Kirata K, Bunai Y, et al. Japanese population DNA typing data for the loci LDLR, GYPA, HBGG, D7S8, and GC. J Forensic Sci 1997;42(5):911–3.
14. Jankowski LB, Budowle B, Swee NT, Pino JA, FreckTootell S, Corey HW, et al. New Jersey Caucasian, African American, and Hispanic population data on the PCR-based loci HLA-DQA1, LDLR, GYPA, HBGG, D7S58, and Gc. J Forensic Sci 1998;43(5):1037–40.
15. Sugiyama E, Honda K, Katsuyama Y, Uchiyama S, Tsuchikane A, Ota M, et al. Allele frequency distribution of the D1S80 (pMCT118) locus polymorphism in the Japanese population by the polymerase chain reaction. Intl J of Legal Med 1993;106(3):111–4.
16. Hedrick PW. Genetics of Populations, Science Books International, Portola Valley, California, 1983;64.

Additional information and reprint requests:
William Klitz, Ph.D.
School of Public Health
University of California
140 Warren Hall
Berkeley, CA 94720-7360